

# On penalized maximum likelihood estimation of approximate factor models \*

Shaoxin Wang<sup>†</sup>, Hu Yang, Chaoli Yao

College of Mathematics and Statistics, Chongqing University,  
Chongqing, 401331, China

August 23, 2016

## Abstract

In this paper, we mainly focus on the estimation of high-dimensional approximate factor model. We rewrite the estimation of error covariance matrix as a new form which shares similar properties as the penalized maximum likelihood covariance estimator given by [Bien and Tibshirani \(2011\)](#). Based on the lagrangian duality, we propose an APG algorithm to give a positive definite estimate of the error covariance matrix. The new algorithm for the estimation of approximate factor model has a desirable non-increasing property. By keeping the error covariance matrix to be positive definite, the efficiency of the new algorithm on estimation and forecasting is investigated via extensive simulations and real data analysis.

*Key words:* Approximate factor model, error covariance matrix, positivity, EM algorithm, APG algorithm

---

\*The work is supported National Natural Science Foundation of China (Grant Nos: 11471059, 11171361)

<sup>†</sup>Corresponding author. Emails: [sxwang@cqu.edu.cn](mailto:sxwang@cqu.edu.cn); [sxwangsd@163.com](mailto:sxwangsd@163.com)(S.Wang), [hy@cqu.edu.cn](mailto:hy@cqu.edu.cn) (H.Yang), [clyao@cqu.edu.cn](mailto:clyao@cqu.edu.cn) (C.Yao)

# 1 Introduction

The factor model finds popularity in psychometrics, economics and many other research fields for its utility in summarizing information in large dataset. Many researchers have investigated different topics related to the factor model, such as its estimation and inference theory (see e.g. [Mulaik, 2009](#); [Bai, 2003](#)), the dynamic factor model and its generalization (see e.g. [Geweke, 1978](#); [Forni et al., 2000](#)), and so on. Another active topic relating to factor model is covariance matrix estimation, whose recent developments can be found in reviews by [Bai and Shi \(2011\)](#), [Fan et al. \(2016\)](#) and the reference therein.

Suppose  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$  are  $p$  dimensional independent and identically distributed (i.i.d.) random vectors. Assume that the mean of  $\mathbf{y}_i$  is  $\boldsymbol{\mu}$  and its covariance matrix is  $\boldsymbol{\Sigma}_{\mathbf{y}}$ . The factor model can be stated as follows

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_i + \mathbf{e}_i, \quad (1.1)$$

where  $\boldsymbol{\Lambda}$  is the  $p \times r$  factor loading matrix, and  $\mathbf{f}_i$  is a column vector of  $r$  components with mean  $E(\mathbf{f}) = \mathbf{0}$  and uncorrelated with the idiosyncratic error  $\mathbf{e}_i$ . Then the covariance matrix  $\boldsymbol{\Sigma}_{\mathbf{y}}$  can be decomposed as

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \boldsymbol{\Lambda}\text{Cov}(\mathbf{f})\boldsymbol{\Lambda}^T + \boldsymbol{\Sigma}_{\mathbf{e}}. \quad (1.2)$$

When the covariance matrix of  $\mathbf{e}_i$ ,  $\text{Cov}(\mathbf{e}_i) = \boldsymbol{\Sigma}_{\mathbf{e}}$ , is diagonal matrix, model (1.1) is called the strict factor model, which has been used by [Ross \(1976\)](#) in studying the arbitrage in market, and showed that when the strict factor model is satisfied, the mean  $\boldsymbol{\mu}$  of asset is approximately linear function of factor loading. [Chamberlain and Rothschild \(1983\)](#) relaxed the diagonal assumption and showed that as long as the eigenvalue of  $\boldsymbol{\Sigma}_{\mathbf{e}}$  is bounded the conclusion still holds. This leads to the approximate factor model, which allows the dependence among the error terms.

The distribution of  $\mathbf{y}_i$  is usually assumed to be nondegenerate, or equivalent to say  $\boldsymbol{\Sigma}_{\mathbf{y}}$  is positive definite. A sufficient condition is that  $\boldsymbol{\Sigma}_{\mathbf{e}}$  is positive definite, we use this assumption throughout the paper. To estimate model (1.1), maximum likelihood (ML) method will be employed to give an simultaneous estimation of  $\boldsymbol{\Lambda}$  and  $\boldsymbol{\Sigma}_{\mathbf{e}}$ . It should be noted that once the factor loadings is given, the

factor can be estimated by weighted least squares. However, the maximum likelihood estimation (MLE) of  $\mathbf{\Lambda}$  is largely depending on the estimation of  $\mathbf{\Sigma_e}$ . In high-dimensional setting,  $p > n$ , the estimation of  $\mathbf{\Sigma_e}$  is difficult due to the fact that estimating too many parameters  $\mathcal{O}(p^2)$  with a relative small sample size. Therefore, additional structure assumption on  $\mathbf{\Sigma_e}$  is usually needed. One typical and widely used is that  $\mathbf{\Sigma_e}$  is conditional sparse, which requires many off-diagonal elements to be zeros or nearly zeros (Bai and Liao, 2016). To get a sparse estimate of  $\mathbf{\Sigma_e}$ , the penalization can be used, such as LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006) and the folded concave penalization including SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) as its special cases.

To estimate  $\mathbf{\Lambda}$ , some identification conditions are needed. We follow the usual restriction (Lawley and Maxwell, 1971; Bai and Liao, 2016) used for MLE of factor model:

$$\text{Cov}(\mathbf{f}) = I_r, \text{ and } \mathbf{\Lambda}^T \mathbf{\Sigma_e}^{-1} \mathbf{\Lambda} \text{ is diagonal,} \quad (1.3)$$

and the elements of  $\mathbf{\Lambda}^T \mathbf{\Sigma_e}^{-1} \mathbf{\Lambda}$  are distinct and arranged in descending order. Under the restriction (1.3), Bai and Liao (2016) proposed the following penalized (quasi) MLE to estimate the parameters

$$(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Sigma_e}}) = \arg \min_{\mathbf{\Lambda}, \mathbf{\Sigma_e}} \log(|\det(\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Sigma_e})|) + \text{tr}((\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Sigma_e})^{-1} \mathbf{S_y}) + P_\lambda(\mathbf{\Sigma_e}), \quad (1.4)$$

which is solved by an majorize-minimize EM algorithm. To be specific, the  $\mathbf{\Lambda}$  is updated by a formula derived from EM algorithm, then  $\mathbf{\Sigma_e}$  updated by majorize-minimize (MM) procedure. We find that due to thresholding operator, the algorithm in high-dimensional setting can not guarantee the  $\hat{\mathbf{\Sigma_e}}$  to be positive definite, even  $\mathbf{\Sigma_y}$ . This motivates our work. Based on the EM algorithm, after updating the  $\mathbf{\Lambda}$  we rewrite the process of updating  $\mathbf{\Sigma_e}$  as a penalized ML covariance matrix estimation problem. Then we propose a new algorithm largely depending on Moreau-Yosida regularization to compute the estimate of  $\mathbf{\Sigma_e}$ . The new algorithm guarantees  $\hat{\mathbf{\Sigma_e}}$  is positive definite. Due to the nonconvex property of the object function, although global convergence can not be guaranteed, we show the proposed algorithm has a nonincreasing property.

The rest of the paper is organized as follows. Section 2 contains some preliminaries and necessary results. The modified algorithm is proposed in the Section 3. All the simulation and a real data analysis are presented in Section 4. The concluding discussion and the proof of main results are given in Section 5 and Appendix correspondingly.

## 2 Penalized estimation of approximate factor model

When we take  $\mathbf{f}$  and  $\mathbf{e}$  as multivariate normal random vectors, and have a random sample of  $n$  observations  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ , then the negative log-likelihood function is given by

$$L(\mathbf{\Lambda}, \mathbf{\Sigma_e}) = \frac{n}{2} \log(|\mathbf{\Sigma_y}|) + \frac{n}{2} \text{tr}(\mathbf{\Sigma_y}^{-1} \mathbf{S_y}) + \frac{n}{2} p \log(2\pi),$$

where  $\mathbf{S_y} = \frac{1}{n} \sum_{j=1}^n (\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^T$  and  $\bar{\mathbf{y}} = \frac{1}{n} \sum_{j=1}^n \mathbf{y}_j$  is the MLE of  $\boldsymbol{\mu}$ . Since a parsimonious estimate of  $\mathbf{\Sigma_e}$  is desired, we put the penalty function  $P_\lambda(\cdot)$  on  $\mathbf{\Sigma_e}$  to penalize the small off-diagonal element of  $\mathbf{\Sigma_e}$  to be zero. Following Bai and Liao (2016), the weighted  $L1$ -penalty is used

$$P_\lambda(\mathbf{\Sigma_e}) = \lambda \|\mathbf{W} \circ \mathbf{\Sigma_e}\|_1,$$

where the symbol  $\circ$  denotes the Hadamard product. This penalty function includes LASSO, adaptive LASSO, and SCAD as its special cases by verifying the elements of matrix  $\mathbf{W}$ . When the serial and cross-sectional dependence between the elements of the error term is allowed, we get the following penalized MLE function

$$(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Sigma_e}}) = \arg \min_{\mathbf{\Lambda}, \mathbf{\Sigma_e}} \log(|\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Sigma_e}|) + \text{tr}((\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Sigma_e})^{-1} \mathbf{S_y}) + P_\lambda(\mathbf{\Sigma_e}), \quad (2.1)$$

where  $(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Sigma_e}})$  are the minimizers of the above penalized MLE problem. It should be noted that equation (2.1) is different from (1.4) for the log part even in our normal distribution setting. By discarding taking the absolute value of the determinant of  $\mathbf{\Sigma_y} = \mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Sigma_e}$ , equation (2.1) guarantees  $\hat{\mathbf{\Sigma_y}} = \hat{\mathbf{\Lambda}} \hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Sigma_e}}$  to be positive definite, which may not be ensured by (1.4). Here we

need to make some clarification on the symbol  $|\cdot|$  to avoid confusion, in (1.4)  $|\cdot|$  means taking the absolute value of the determinant of matrix  $\Sigma_{\mathbf{y}} = \Lambda\Lambda^T + \Sigma_{\mathbf{e}}$ , whereas in the rest of the paper  $|\cdot|$  means taking the determinant of a matrix unless otherwise specified.

## 2.1 The Moreau-Yosida regularization

Here, we need to present some definition and results on the Moreau-Yosida regulation, which provides useful tools in designing an accelerated proximal gradient (APG) algorithm.

Let  $f$  be a closed proper convex function. The Moreau-Yosida regularization of  $f$  associated with a given parameter  $\rho$  is defined as

$$\varphi_f^\rho(x) = \min_{y \in \mathcal{X}} \left\{ f(y) + \frac{1}{2\rho} \|y - x\|^2 \right\}, \quad (2.2)$$

where  $x \in \mathcal{X}$ , the domain of  $f$  and equipped with the norm  $\|\cdot\|$ . The unique minimizer of (2.2), denoted by  $P_f^\rho(x)$ , is called the proximal point mapping associated with  $f$ . Cui et al. (2016) presented the following proposition building some important connections between  $\varphi_f^\rho(x)$  and  $P_f^\rho(x)$ , whose proof can be found in (Hiriart-Urruty and Lemaréchal, 1993, p.320).

**Proposition 1.** Let  $f : \mathcal{X} \rightarrow (-\infty, +\infty]$  be a closed proper convex function,  $\varphi_f^\rho$  is the Moreau-Yosida regularization of  $f$ , and  $P_f^\rho$  be the associated proximal point mapping. Then the following results hold.

1.  $\varphi_f^\rho$  is continuously differentiable with gradient given by

$$\nabla \varphi_f^\rho(x) = \frac{1}{\rho}(x - P_f^\rho(x)).$$

Furthermore,  $\nabla \varphi_f^\rho$  is continuously Lipschitz continuous with modulus  $1/\rho$ .

2. For any  $x_1, x_2 \in \mathcal{X}$ , one has

$$\langle P_f^\rho(x_1) - P_f^\rho(x_2), x_1 - x_2 \rangle \geq \|P_f^\rho(x_1) - P_f^\rho(x_2)\|^2,$$

which implies the mapping  $P_f^\rho(\cdot)$  is globally continuous with modulus 1 by the Cauchy-Schwarz inequality.

The Proposition 1 is useful in coping with the constrained optimization problem involving nonsmooth function, and the Moreau-Yosida regularization of any closed proper convex function is continuously differentiable.

### 3 Algorithm

The EM-algorithm has been widely used in finding the MLE of factor model (see e.g. [Rubin and Thayer, 1982](#); [Hirose and Yamamoto, 2015](#)). In this section, we derive an EM based iterative procedure for computing the penalized estimator. Let  $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_n]$  be the missing data and  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]$  the observed one, then given the current estimates of factor loading matrix and error covariance matrix, denoted by  $\mathbf{\Lambda}_{old}$  and  $\mathbf{\Sigma}_{eold}$ , the conditional expectation of the penalized log-likelihood function of complete data given  $\mathbf{Y}$  is

$$\begin{aligned} E(L_c(\mathbf{\Lambda}, \mathbf{\Sigma}_e) | \mathbf{Y}) = & -\frac{n}{2} \log(|\mathbf{\Sigma}_e|) - \frac{n}{2} \text{tr}(\mathbf{\Sigma}_e^{-1} \mathbf{S}_y) + n \text{tr}(\mathbf{\Lambda}^T \mathbf{\Sigma}_e^{-1} \mathbf{S}_y \mathbf{\Gamma}_{old}) - \frac{n}{2} P_\lambda(\mathbf{\Sigma}_e) \\ & - \frac{n}{2} \text{tr}((\mathbf{\Lambda}^T \mathbf{\Sigma}_e^{-1} \mathbf{\Lambda} + I_r)(\mathbf{\Gamma}_{old}^T \mathbf{S}_y \mathbf{\Gamma}_{old} + \mathbf{\Omega}_{old})) + \text{const.} \quad , \end{aligned} \quad (3.1)$$

where  $\mathbf{\Gamma}_{old} = \mathbf{\Sigma}_{yold}^{-1} \mathbf{\Lambda}_{old}$  with  $\mathbf{\Sigma}_{yold} = \mathbf{\Lambda}_{old} \mathbf{\Lambda}_{old}^T + \mathbf{\Sigma}_{eold}$ , and  $\mathbf{\Omega}_{old} = I_r - \mathbf{\Lambda}_{old}^T \mathbf{\Gamma}_{old}$ , the derivation of (3.1) is given in Section A. Therefore, we calculate  $\mathbf{\Lambda}_{new}$  and  $\mathbf{\Sigma}_{enew}$  by solving the following penalized minimization problem

$$\begin{aligned} (\mathbf{\Lambda}_{new}, \mathbf{\Sigma}_{enew}) = & \arg \min_{\mathbf{\Lambda}, \mathbf{\Sigma}_e} \log(|\mathbf{\Sigma}_e|) + \text{tr}(\mathbf{\Sigma}_e^{-1} \mathbf{\Lambda} \mathbf{\Omega}_{old} \mathbf{\Lambda}^T) \\ & + \text{tr}(\mathbf{\Sigma}_e^{-1} (I_p - \mathbf{\Lambda} \mathbf{\Gamma}_{old}^T) \mathbf{S}_y (I_p - \mathbf{\Lambda} \mathbf{\Gamma}_{old}^T)^T) + P_\lambda(\mathbf{\Sigma}_e). \end{aligned} \quad (3.2)$$

Treating  $\mathbf{\Sigma}_e$  as a constant, we take the derivative of (3.2) and update  $\mathbf{\Lambda}$  as follows

$$\mathbf{\Lambda}_{new} = \mathbf{S}_y \mathbf{\Gamma}_{old} (\mathbf{\Omega}_{old} + \mathbf{\Gamma}_{old}^T \mathbf{S}_y \mathbf{\Gamma}_{old})^{-1}, \quad (3.3)$$

then substitute  $\Lambda_{new}$  into equation (3.2), and update  $\Sigma_e$  by

$$\Sigma_{enew} = \arg \min_{\Sigma_e} \log(|\Sigma_e|) + \text{tr}(\Sigma_e^{-1}M) + P_\lambda(\Sigma_e), \quad (3.4)$$

where

$$M = (I_p - \Lambda_{new}\Gamma_{old}^T)\mathbf{S}_y(I_p - \Lambda_{new}\Gamma_{old}^T)^T + \Lambda_{new}\Omega_{old}\Lambda_{new}^T. \quad (3.5)$$

Here, we need to claim that the  $\Lambda$ s in  $\Gamma$  and  $\Omega$  are also updated. Since  $\Sigma_{eold}$  is still unchanged, we here just write  $\Gamma_{old}$  and  $\Omega_{old}$  for the simplicity of presentation. We note that equation (3.4) has the same form as the penalized MLE of covariance matrix discussed by [Bien and Tibshirani \(2011\)](#), except for the matrix  $M$ . But from equation (3.5), it can be easily checked that when  $\mathbf{S}_y$  is positive definite,  $M$  is also positive definite. By the Proposition 1 in ([Bien and Tibshirani, 2011](#)), we may conclude that the matrix  $M$  has the same property and influence as  $\mathbf{S}_y$  on the solution of (3.4). So we adopt the MM procedure used in ([Bien and Tibshirani, 2011](#)) to calculate an approximate solution of problem (3.4). Since  $\log(|\Sigma_e|)$  is concave, by the MM procedure and the proximal gradient method, we can approximate the original problem (3.4) by

$$\Sigma_{enew} = \arg \min_{\Sigma_e \succeq \delta I_p} \frac{1}{2t} \|\Sigma_e - \mathcal{M}_n\|_F^2 + P_\lambda(\Sigma_e), \quad (3.6)$$

where

$$\mathcal{M}_n = \Sigma_{eold} - t(\Sigma_{eold}^{-1} - (\Sigma_{eold}^{-1}M\Sigma_{eold}^{-1})),$$

and  $t$  is the depth of projection, details on the derivation of (3.6) can be found in ([Bien and Tibshirani, 2011](#)). We need to clarify some motivation behind equation (3.6). The  $\log|\Sigma_e|$  part in (3.4) guarantees  $\Sigma_{enew}$  to be positive definite, and from ([Bien and Tibshirani, 2011](#)) when  $\mathbf{S}_y$  is positive definite, the minimum eigenvalue of  $\Sigma_{enew}$  is larger than some  $\delta > 0$ . However, when  $p > n$ ,  $\mathbf{S}_y$  (or  $M$ ) is semipositive definite. Although [Bien and Tibshirani \(2011\)](#) suggested an augmenting data set procedure to keep  $\mathbf{S}_y$  positive definite, this still can not avoid the minimum

eigenvalue of  $\Sigma_e$  to be negative due to the thresholding procedure. So in high-dimensional setting we might be slightly reckless to give an lower bound  $\delta$ , e.g.  $\delta = 10^{-5}$  used in our simulation, of the minimum eigenvalue of  $\Sigma_e$ , which forms the constraint  $\Sigma_e \succeq \delta I_p$  in (3.6).

To solve problem (3.6), many methods, like alternating direction method of multipliers (ADMM, see e.g. Boyd et al., 2011), can be used. However when ADMM is used, a penalty parameter in the augmented lagrange function need the user to choose, which has no influence on theoretical convergence results (Xue et al., 2012), but can give nonnegligible impact on numerical performance (Ma et al., 2013). Here, we follow Cui et al. (2016) and solve problem (3.6) by applying the APG algorithm to its lagrangian dual problem, which leads to the solution of problem (3.6). The lagrange function of problem (3.6) is

$$L(\Sigma_e, Z) = \frac{1}{2t} \|\Sigma_e - \mathcal{M}_n\|_F^2 + P_\lambda(\Sigma_e) - \langle Z, \Sigma_e - \delta I_p \rangle, \quad (3.7)$$

where  $Z$  is the lagrange multiplier. Its lagrange dual problem is given by

$$\max_{Z \succeq 0} \min_{\Sigma_e} L(\Sigma_e, Z). \quad (3.8)$$

Let  $X(Z) = \mathcal{M}_n + tZ$ , we define and solve the following problem

$$\begin{aligned} g(Z) &:= - \min_{\Sigma_e} \{L(\Sigma_e, Z)\} \\ &= - \inf_{\Sigma_e} \left\{ \frac{1}{2t} \|\Sigma_e - X(Z)\|_F^2 + P_\lambda(\Sigma_e) \right\} \\ &\quad - \frac{1}{2t} \|\delta I_p - \mathcal{M}_n\|_F^2 + \frac{1}{2t} \|X(Z) - \delta I_p\|_F^2. \end{aligned} \quad (3.9)$$

With  $P_\lambda(\Sigma_e) = \lambda \|W \circ \Sigma_e\|_1$ , we get

$$g(Z) = -\frac{1}{2t} \left\| \widehat{\Sigma}_e - X(Z) \right\|_F^2 - \frac{1}{2t} \|\delta I_p - \mathcal{M}_n\|_F^2 + \frac{1}{2t} \|X(Z) - \delta I_p\|_F^2 - P_\lambda(\widehat{\Sigma}_e),$$



where

$$\begin{aligned}\hat{\Sigma}_e &= \mathcal{S}(X(Z), t\lambda W) \\ &= \text{sign}(X(Z)_{ij}) \max\{|X(Z)_{ij}| - t\lambda W_{ij}, 0\}\end{aligned}\tag{3.10}$$

is the minimizer of the first term in (3.9), and  $\mathcal{S}(\cdot, \cdot)$  is the elementwise soft-thresholding operator. By Proposition 1, the function  $g(Z)$  is continuously differentiable, whose gradient is

$$\nabla g(Z) = \mathcal{S}(X(Z), t\lambda W) - \delta I_p.$$

Meanwhile, the lagrange dual problem (3.8) can be written as

$$\min_Z f(Z) \equiv g(Z) + \delta_{psd}(Z),\tag{3.11}$$

where  $\delta_{psd}(\cdot)$  is the indicator function of the cone of positive semidefinite matrices. Beck and Teboulle (2009) proposed an APG algorithm to deal with (3.11), and proved a complexity result of  $\mathcal{O}(1/k^2)$ . Then, from Part 2 of Proposition 1,  $\nabla g(Z)$  is globally Lipschitz continuous, and according to APG algorithm we can solve (3.11) by an iterative procedure of minimizing its quadratic approximation as follows

$$Y_{new} = \arg \min_{Z \succeq 0} g(Z_{old}) + \frac{1}{2} \|Z - Z_{old}\|_F^2 + \langle \nabla g(Z_{old}), Z - Z_{old} \rangle.\tag{3.12}$$

It is well known that the solution to problem (3.12) is given by

$$Y_{new} = \mathcal{P}_+(Z_{old} - \nabla g(Z_{old})),\tag{3.13}$$

$\mathcal{P}_+(\cdot)$  is the projection operator onto the cone of positive semidefinite matrices. For a symmetric matrix  $X$ , its projection onto the cone of positive semidefinite matrices can be computed through its eigen-decomposition  $X = \sum_{i=1}^p \lambda_i v_i v_i^T$ , then  $\mathcal{P}_+(X) = \sum_{i=1}^p \max\{\lambda_i, 0\} v_i v_i^T$ . The above described

procedure forms the following APG algorithm for updating  $\Sigma_e$ .

---

**Algorithm 1 Update  $\Sigma_e$  by APG algorithm**

---

Given  $\Sigma_{ek}$ , the updated estimate  $\Lambda_{k+1}$ , and  $t_k = 1$ . Set  $k := 1$ . Iterate until convergence:

**Step 1** Compute  $M$  by (3.5) and  $\mathcal{M}_n$  in (3.7).

**Step 2** Compute  $\mathcal{S}(X(Z_k), t\lambda W)$  by (3.10),  $\nabla g(Z_k)$ , and  $Y_{k+1} = \mathcal{P}_+(Z_k - \nabla g(Z_k))$ .

**Step 3** Compute  $t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$ .

**Step 4** Compute  $Z_{k+1} = Y_{k+1} + \frac{t_k - 1}{t_{k+1}}(Y_{k+1} + Y_k)$

---

Cui et al. (2016) presented the following nonasymptotic global rate of convergence for APG algorithm, which shows the  $\mathcal{O}(1/k^2)$  complexity result of APG algorithm. Its proof rests on Part 2 of Proposition 1 and Theorem 4.4 of Beck and Teboulle (2009), which is a straight extension from vectors to matrices.

**Theorem 1.** *Assume  $f$  is defined in (3.11) and  $\{Z_k\}$  are generated by the **APG** algorithm. Then for any optimal solution  $Z^*$  of  $\min_Z f(Z)$ , we have*

$$f(Z_k) - f(Z^*) \leq \frac{2\|Z_0 - Z^*\|_F^2}{(k+1)^2}, \quad \forall k \geq 1,$$

where  $Z_0$  is the initial value.

Then, we can conclude this part by the following algorithm giving simultaneous estimates of  $\Lambda$  and  $\Sigma_e$ , and a theorem guaranteeing the algorithm is nonincreasing. The proof of Theorem 2 is given in Section B.

---

**Algorithm 2 EM plus APG algorithm**

---

Given the consistent estimates  $\Lambda_1$ ,  $\Sigma_{e1}$ ,  $t_1 = 1$ . Set  $k := 1$ . Iterate until convergence:

**Step 1** Update  $\Lambda$  by (3.3).

**Step 2** Update  $\Sigma_e$  by APG algorithm.

---

**Theorem 2.** *If we solve the problem (2.1) by **EM plus APG** algorithm, then the object value of each iteration is nonincreasing.*

**Remark 1.** All the discussions given above are of particular interest for high-dimensional setting. When  $p \leq n$ , [Bien and Tibshirani \(2011\)](#) showed that it is often the case that  $\widehat{\Sigma}_{\mathbf{e}}$  can be computed by solving (3.6) without the constraint  $\Sigma_{\mathbf{e}} \succeq \delta I_p$ , and we get

$$\widehat{\Sigma}_{\mathbf{e}} = \mathcal{S}(\mathcal{M}_n, t\lambda W). \quad (3.14)$$

In this situation, the proposed APG algorithm serves as an intermediate step to preserve the positive definiteness of  $\widehat{\Sigma}_{\mathbf{e}}$ . Whereas, [Bien and Tibshirani \(2011\)](#) employed ADMM to ensure  $\widehat{\Sigma}_{\mathbf{e}}$  to be positive definite. The difference between ADMM and APG algorithms has been discussed in ([Cui et al., 2016](#)). We will not repeat it again.

## 4 Empirical study

It should be noted that our main aim is to improve the efficiency of the estimation for  $\mathbf{\Lambda}$  and  $\mathbf{f}$  via preserving the positive definiteness of error covariance matrix  $\Sigma_{\mathbf{e}}$ . We use the  $K$ -fold cross validation to choose the tuning parameter  $\lambda$ , which has been widely used in covariance matrix estimation and finds its theoretical support in ([Bickel and Levina, 2008](#)). Let  $\mathcal{A}$  be the index set of  $n$  observations, and  $\mathbf{S}_{\mathbf{y}, \mathcal{A}} = |\mathcal{A}|^{-1} \sum_{t \in \mathcal{A}} (\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})^T$  the sample covariance matrix given by validation data. The symbol  $|\mathcal{A}|$  denotes the cardinality of set  $\mathcal{A}$ . Let  $\widehat{\mathbf{\Lambda}}(\mathcal{A}^c, \lambda)$  and  $\widehat{\Sigma}_{\mathbf{e}}(\mathcal{A}^c, \lambda)$  be the estimated loading matrix and error covariance matrix by the training data in  $\mathcal{A}^c$ , the compliment of  $\mathcal{A}$ , with tuning parameter  $\lambda$ . Then we partition the data in to  $K$  subsets, denoted by its index sets  $\mathcal{A}_1, \dots, \mathcal{A}_K$ , and choose the  $\lambda_{cv}$  as follows

$$\lambda_{cv} = \arg \min_{\lambda} \frac{1}{K} \sum_{k=1}^K L \left( \widehat{\mathbf{\Lambda}}(\mathcal{A}_k^c, \lambda), \widehat{\Sigma}_{\mathbf{e}}(\mathcal{A}_k^c, \lambda), \mathbf{S}_{\mathbf{y}, \mathcal{A}_k} \right),$$

where  $L(\widehat{\mathbf{\Lambda}}, \widehat{\Sigma}_{\mathbf{e}}, \mathbf{S}_{\mathbf{y}}) = \frac{1}{n} \log \left( |\widehat{\mathbf{\Lambda}} \widehat{\mathbf{\Lambda}}^T + \widehat{\Sigma}_{\mathbf{e}}| \right) + \frac{1}{n} \text{tr}(\mathbf{S}_{\mathbf{y}}(\widehat{\mathbf{\Lambda}} \widehat{\mathbf{\Lambda}}^T + \widehat{\Sigma}_{\mathbf{e}})^{-1})$ .

## 4.1 Simulation

In our simulation, we take the synthetic example given by [Bai and Liao \(2016\)](#) with only some modification on the error covariance matrix  $\Sigma_e$  to check the efficiency of our new algorithm. The two different error covariance matrices given below satisfy the conditional sparsity condition given in [\(Bai and Liao, 2016\)](#).

**Model 1.** (*Banded Matrix*) Let  $\{\alpha_{ij}\}_{i \leq p, j \leq n}$  be generated from standard normal distribution  $\mathcal{N}(0, 1)$ , and then

$$\begin{aligned} e_{1,j} &= \alpha_{1,j}, \quad e_{2,j} = \alpha_{2,j} + a_1 \alpha_{1,j}, \\ e_{3,j} &= \alpha_{3,j} + a_2 \alpha_{2,j} + b_1 \alpha_{1,j}, \\ e_{i+1,j} &= \alpha_{i+1,j} + a_i \alpha_{i,j} + b_{i-1} \alpha_{i-1,j} + c_{i-2} \alpha_{i-2,j}, \end{aligned}$$

where  $\{a_i, b_i, c_i\}_{i \leq p}$  are independently from  $0.7 \times \mathcal{N}(0, 1)$ . Let the two factors  $\{f_{1j}, f_{2j}\}$  be i.i.d.  $\mathcal{N}(0, 1)$ , and the elements of  $\mathbf{\Lambda}$  be uniform on  $[0, 1]$ .

**Model 2.** (*Approximately Sparse Matrix*) Let  $\Sigma_e = \alpha I_p + M$ , where  $M = [m_{ij}]$  with  $m_{ij} = 0.5^{|i-j|}$ , and  $\alpha$  is used to control the condition number of  $\Sigma_e$  equal to  $p$ .

Since we want to investigate the influence of the positivity of  $\Sigma_e$  on the estimation of  $\mathbf{\Lambda}$  and  $\mathbf{f}$ , we use three different methods: (1) MMEM: the majorize-minimize EM algorithm proposed by [Bai and Liao \(2016\)](#); (2) EMAPG: our proposed algorithm; (3) EPC: a two-step algorithm based on generalized principal component method ([Choi, 2012](#)) to estimate the loading matrix and factor, and principal orthogonal complement thresholding (POET) method by [Fan et al. \(2013\)](#) to estimate the error covariance matrix. When using POET estimator, we follow [Bai and Liao \(2016\)](#) to choose the thresholding parameter  $C = 0.7, 0.5$ . We repeat the experiment 100 times. To compare the performance of these three algorithms, the canonical correlation analysis for  $\mathbf{\Lambda}$  and  $\mathbf{f}$  is used due to its invariant property under invertible transformation, and the higher value implies better performance. For the estimation of  $\Sigma_e$ , we use the root-mean-square error (RMSE),  $\|\hat{\Sigma}_e - \Sigma_e\|_F / p$ . We also report the times of  $\Sigma_e | \Sigma_y$  to be non-positive definite (NPD) in 100

replications. All the simulation results for Model 1 and Model 2 are reported in Table 1 and Table 2, correspondingly.

From both Table 1 and Table 2, we can find that except for some special case ( $n = 100, p = 50$ ) our algorithm outperforms the MMEM algorithm, which implies that preserving the positive definiteness of  $\Sigma_e$  can improve the estimation of  $\Lambda$  and  $\mathbf{f}$ , and the estimation of  $\Sigma_e$  has non-negligible impact on the estimation of  $\Lambda$  and  $\mathbf{f}$ . Specifically, when the percentage of nonpositivity of  $\hat{\Sigma}_e$  (or  $\hat{\Sigma}_y$ ) is higher, the superiority of our algorithm becomes more obvious. Moreover, our numerical result also shows that the MMEM algorithm performs not stable in giving the estimation of  $\Sigma_e$  for its higher variance of RMSE. From Table 2, we note that the EPC algorithm gives best performance when  $\hat{\Sigma}_e$  is positive definite. This coincides the conclusion given in (Bai and Liao, 2016) that it is hard to see whether MMEM(EMAPG) or EPC dominates the other. In general, we find that preserving the positive definiteness of the error covariance matrix can improve the estimation of loading matrix and factor for both the one step penalized MLE and the two step EPC methods. Another issue should be addressed is CUP time of the algorithms, all the simulations are performed in Matlab on a PC AMD Athlon(tm) II X2 215 processor, 2.7Ghz. We find that for Model 1 the CPU time of EMAPG is usually longer than that of MMEM due to guaranteeing  $\hat{\Sigma}_e$  positive definite, but no more than 10 times. However, for Model 2 the things change dramatically. The CPU time of EMAPG can be much and even about 100 times longer than that of MMEM, and then the reward of high CPU time may be its best performance in the simulation.

## 4.2 Forecasting simulation

It has been shown by Bai and Liao (2016) the ML based and PC based methods give different procedures for estimating  $\Sigma_e$ . The simulation in (Bai and Liao, 2016) demonstrates that considering the cross-sectional correlations leads to a considerable improvement of the estimation of  $\Lambda$  and  $\mathbf{f}$ , and results in better performance of forecasting. In our simulation, we have shown the estimation of  $\Lambda$  and  $\mathbf{f}$  can be further improved by preserving  $\hat{\Sigma}_e$  to be positive definite. So we want to check the influence of preserving  $\hat{\Sigma}_e$  positive definite on the forecasting. Confining ourself to the same framework of ML based methods, we only study the following three methods to investigate the

Table 1: Comparison of three methods for Model 1.

Model 1	$n$	$p$	MMEM	EMAPG	EPC	
			CV	CV	$C = 0.7$	$C = 0.5$
Loadings	50	50	0.3510	0.3579	0.2869	0.2658
		100	0.4740	0.4937	0.4475	0.3829
		150	0.6028	0.6534	0.6043	0.4891
	100	50	0.3335	0.3315	0.2916	0.2681
		100	0.6333	0.6774	0.5408	0.4131
		150	0.7693	0.8074	0.7700	0.5909
Factors	50	50	0.3523	0.3721	0.3108	0.2838
		100	0.5415	0.5759	0.5142	0.3825
		150	0.6811	0.7281	0.7120	0.5133
	100	50	0.3265	0.3265	0.2938	0.2664
		100	0.6827	0.7432	0.6135	0.4052
		150	0.8697	0.9019	0.8622	0.5300
NPD	50	50	14% 1%	0% 0%	0% 0%	23% 23%
		100	45% 36%	0% 0%	0% 0%	91% 91%
		150	69% 64%	0% 0%	0% 0%	100% 100%
	100	50	15% 5%	0% 0%	0% 0%	27% 14%
		100	14% 6%	0% 0%	0% 0%	78% 78%
		150	23% 18%	0% 0%	7% 7%	100% 100%
RMSE	50	50	0.4197(0.4857)	0.2873(0.0001)	0.2499(0.0001)	0.2317(0.0001)
		100	0.6208(3.9193)	0.2003(0.0000)	0.1709(0.0000)	0.1680(0.0000)
		150	6.4191(172.2575)	0.1754(0.0000)	0.1354(0.0000)	0.1386(0.0000)
	100	50	1.0733(34.2446)	0.3148(0.0013)	0.2676(0.0001)	0.2523(0.0001)
		100	1.1189(72.3207)	0.1735(0.0001)	0.1240(0.0000)	0.1119(0.0000)
		150	1.1445(28.6522)	0.1556(0.0000)	0.1050(0.0000)	0.1011(0.0000)

influence of  $\Sigma_e$  on forecasting: (1) the heteroscedastic ML (HML) estimator without considering the cross-sectional dependence; (2) the MMEM algorithm; (3) the EMAPG algorithm.

We consider the following synthetic two factors time series model:

$$x_{t+1} = \beta^T \mathbf{f}_t + \epsilon_t, \quad \mathbf{f}_t = \rho \mathbf{f}_{t-1} + \nu_t,$$

where  $\beta = \begin{bmatrix} 2, & 3 \end{bmatrix}^T$ ,  $\rho = 0.5$ ,  $\epsilon_t \sim \mathcal{N}(0, 1)$ ,  $\nu_t \sim \mathcal{N}(0, I_2)$ , and  $\mathbf{f}_t$  is the unknown factor and can be estimated from the following factor model:

$$\mathbf{y}_t = \Lambda \mathbf{f}_t + \mathbf{e}_t.$$

In our simulation we set  $\mathbf{e}_t$  to have the covariance structure as Model 2, and  $n = 50, 100$  be the sample size. We first generate  $m + n$  observations, then use the first  $n$  observations to estimate the

Table 2: Comparison of three methods for Model 2.

Model 2	$n$	$p$	MMEM	EMAPG	EPC	
			CV	CV	$C = 0.7$	$C = 0.5$
Loadings	50	50	0.7362	0.7698	0.7926	0.7587
		100	0.8104	0.8743	0.8829	0.7512
		150	0.8608	0.8953	0.8714	0.7894
	100	50	0.8863	0.9337	0.9320	0.8819
		100	0.8725	0.9444	0.8707	0.8856
		150	0.9048	0.9511	0.7773	0.8209
Factors	50	50	0.6706	0.7699	0.8165	0.6864
		100	0.7339	0.9375	0.9654	0.6154
		150	0.7197	0.9456	0.9021	0.6676
	100	50	0.7631	0.9233	0.9290	0.6553
		100	0.7339	0.9534	0.6648	0.6620
		150	0.7368	0.9777	0.5121	0.5964
NPD	50	50	95% 94%	0% 0%	0% 0%	100% 100%
		100	100% 100%	0% 0%	0% 0%	100% 100%
		150	99% 99%	0% 0%	76% 76%	100% 100%
	100	50	95% 93%	0% 0%	5% 2%	100% 100%
		100	100% 99%	0% 0%	100% 100%	100% 100%
		150	100% 100%	0% 0%	100% 100%	100% 100%
RMSE	50	50	14.9604(317.0625)	0.1112(0.0000)	0.0851(0.0000)	0.0768(0.0000)
		100	144.0940(4.2692e5)	0.0784(0.0000)	0.0509(0.0000)	0.0476(0.0000)
		150	217.1154(3.8716e6)	0.0653(0.0001)	0.0401(0.0000)	0.0396(0.0000)
	100	50	30.8220(17.7594)	0.1090(0.0000)	0.0692(0.0000)	0.0612(0.0000)
		100	68.6753(295.5109)	0.0770(0.0000)	0.0401(0.0000)	0.0369(0.0000)
		150	48.5703(4.6827e4)	0.0628(0.0000)	0.0305(0.0000)	0.0293(0.0000)

factor and  $\beta$ , and forecast  $x_{n+1}$ . The process is repeated  $m$  times. To be specific, let  $t = 0, \dots, m-1$  and we first estimate the factor by the data  $\{\mathbf{y}_i\}_{i=t+1}^n$  and get  $\{\hat{\mathbf{f}}_i\}_{i=t+1}^n$ , then we get the estimator  $\hat{\beta}_{t+n}$  by regressing  $\{x_i\}_{i=t+2}^n$  onto  $\{\hat{\mathbf{f}}_i\}_{i=t+1}^{n-1}$ . In the end the forecast of  $x_{t+n+1}$  is  $\hat{x}_{t+n+1} = \hat{\beta}_{t+n}^T \hat{\mathbf{f}}_{t+n}$ , and the forecasting error is  $(x_{t+n+1} - \hat{x}_{t+n+1})^2$ . In order to give quantitative measurement of the performance, we use PC estimator as benchmark and compute the mean squared out-of-sample forecasting error:

$$MSE = \frac{1}{m} \sum_{t=0}^{m-1} (x_{t+T+1}^h - \hat{x}_{t+T+1}^h)^2.$$

For different methods, the relative MSEs to PC method are reported in Table 3. From Table 3, we can see that considering the correlation between idiosyncratic error term can improve the forecast. We also note that the MSEs of EMAPG algorithm usually smaller than the other two methods, which implies that preserving the positive definiteness of  $\Sigma_e$  can improve forecasting.

Table 3: Forecasting error of synthetic data with  $m = 50$ .

		HML	PML	EMAPG
$p = 50$	$n = 50$	0.8411	0.7683	0.6803
	$n = 100$	0.9291	0.9241	0.8738
	$n = 150$	0.9583	0.9595	0.8993
$p = 100$	$n = 50$	0.8558	0.7881	0.7549
	$n = 100$	0.9441	1.0515	0.9851
	$n = 150$	0.9965	1.0085	0.9992

### 4.3 Macroeconomic times series data

We consider the diffusion index forecasting model given by [Stock and Watson \(2002\)](#) and its general form is given by

$$x_{t+h}^h = a_h + \beta_h^T \mathbf{f}_t + \sum_{i=1}^l \gamma_{ih} x_{t+1-i} + \epsilon_{t+h}^h,$$

$$\mathbf{y}_t = \mathbf{\Lambda} \mathbf{f}_t + \mathbf{e}_t,$$

where  $h$  is the forecast horizon,  $l$  is the number of lags for  $x_t$ , and the goal is to forecast  $x_{t+h}^h = \frac{1}{h} \sum_{i=1}^h x_{t+i}$  the  $h$ -step-ahead variable. The dataset of real time macroeconomic time series of the United States is used, which has been analyzed by [Ludvigson and Ng \(2011\)](#) and identified with 8 factors by information criterion. To perform the forecasting, we follow the procedure of forecasting used in Section 4.2 and report the relative MSEs to PC method for the case with 3 lags and 8 factors in Table 4. From Table 4, for  $h = 1$ , the sort term forecast, all these three method has

Table 4: Forecasting error of real data analysis.

	HML	PML	EMAPG
$h = 1$	1.0682	0.9485	1.0086
$h = 12$	1.0522	0.6508	0.7431
$h = 18$	0.9387	0.5967	0.5685
$h = 24$	0.8748	0.6401	0.5381

no very large difference on forecasting. Without considering the potential change of data, taking the cross-sectional dependence into account makes MMEM and EMAPG having smaller MSEs compared with HML method. In addition, preserving  $\hat{\Sigma}_{\mathbf{e}}$  positive definite gives smaller MSE of



EMAPG than MMEMs for the cases  $h = 18$  and  $h = 24$ , but MMEM gives the best performance when  $h = 12$ .

## 5 Concluding discussion

In this paper, we reformulate the penalized maximum likelihood estimation of approximate factor model and the main contribution is we rewrite the estimation of error covariance matrix as a variant of penalized maximum likelihood estimation of covariance matrix. The new form (3.4) shares similar properties with the covariance matrix estimator given by [Bien and Tibshirani \(2011\)](#). We also propose an APG algorithm to give a positive definite estimate of error covariance matrix, which has the comparable performance with ADMM but needs less user-chosen parameter.

From our empirical study, especially Model 2 in Section 4.1, the main observation is that the estimation of approximate factor model can be significantly improved by keeping error covariance matrix positive definite. In the forecasting simulation, keeping the error covariance matrix to be positive definite usually gives small MSEs in our simulations.

Another issue should be addressed is that our new algorithm can produce good estimates of  $\Lambda$  and  $\Sigma_e$ . With equation (1.2), the algorithm can be directly applied to estimate the covariance matrix  $\Sigma_y$  without any difficulty.

## Acknowledgements

The authors would like to appreciate Dr. Yuan Liao for sharing the Matlab codes used in ([Bai and Liao, 2016](#)).

## Appendix

### A Derivation of (3.1)

*Proof.* Let  $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n]$  be the missing data,  $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$  the observed one. Then the log likelihood function of complete data is given by

$$\begin{aligned} L_c(\mathbf{\Lambda}, \mathbf{\Sigma_e}) = & -\frac{n}{2} \log(|\mathbf{\Sigma_e}|) - \frac{1}{2} \sum_{i=1}^n \mathbf{f}_i^T \mathbf{f}_i \\ & - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu} - \mathbf{\Lambda} \mathbf{f}_i)^T \mathbf{\Sigma_e}^{-1} (\mathbf{y}_i - \boldsymbol{\mu} - \mathbf{\Lambda} \mathbf{f}_i) + \text{const.} \end{aligned} \quad (\text{A.1})$$

Since  $\bar{\mathbf{y}}$  is the MLE of  $\boldsymbol{\mu}$  and the conditional distribution of  $\mathbf{f}_i$  given  $\mathbf{y}_i$  is

$$\mathbf{f}_i \sim \mathcal{N}(\mathbf{\Gamma}_{old}^T (\mathbf{y}_i - \bar{\mathbf{y}}), \mathbf{\Omega}_{old}),$$

with  $\mathbf{\Gamma}_{old} = \mathbf{\Sigma}_{\mathbf{y}_{old}}^{-1} \mathbf{\Lambda}_{old}$  and  $\mathbf{\Omega}_{old} = I_r - \mathbf{\Lambda}_{old}^T \mathbf{\Gamma}_{old}$ , the conditional expectation of  $L_c(\mathbf{\Lambda}, \mathbf{\Sigma_e})$  given  $\mathbf{Y}$  is

$$\begin{aligned} \text{E}(L_c(\mathbf{\Lambda}, \mathbf{\Sigma_e}) | \mathbf{Y}) = & -\frac{n}{2} \log(|\mathbf{\Sigma_e}|) - \frac{n}{2} \text{tr}(\mathbf{\Sigma_e}^{-1} \mathbf{S}_{\mathbf{y}}) \\ & - \frac{1}{2} \text{E} \left( \sum_{i=1}^n \mathbf{f}_i^T (\mathbf{\Lambda}^T \mathbf{\Sigma_e}^{-1} \mathbf{\Lambda} + I_r) \mathbf{f}_i | \mathbf{Y} \right) \\ & + \text{E} \left( \sum_{j=1}^n \mathbf{f}_j^T \mathbf{\Lambda}^T \mathbf{\Sigma_e}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}) | \mathbf{Y} \right) + \text{const.} \end{aligned} \quad (\text{A.2})$$

With a little algebra, we get

$$\text{E} \left( \sum_{j=1}^n \mathbf{f}_j \mathbf{f}_j^T | \mathbf{Y} \right) = n(\mathbf{\Gamma}_{old}^T \mathbf{S}_{\mathbf{y}} \mathbf{\Gamma}_{old} + \mathbf{\Omega}_{old}), \quad (\text{A.3})$$

$$\text{E} \left( \sum_{j=1}^n (\mathbf{y}_i - \bar{\mathbf{y}}) \mathbf{f}_j^T | \mathbf{Y} \right) = n \mathbf{S}_{\mathbf{y}} \mathbf{\Gamma}_{old}. \quad (\text{A.4})$$

Substituting (A.3) and (A.4) into (A.2) and adding the penalized term give (3.1).  $\square$

## B Proof of Theorem 2

*Proof.* For simplicity of presentation, we set  $f$  be the generic density function. Let  $f(\mathbf{y}|\mathbf{\Lambda}, \mathbf{\Sigma_e})$  be the density function, then we use the following penalized maximum likelihood function

$$L_p(\mathbf{\Lambda}, \mathbf{\Sigma_e}) = \log(f(\mathbf{Y}|\mathbf{\Lambda}, \mathbf{\Sigma_e})) - \frac{n}{2}P_\lambda(\mathbf{\Sigma_e}), \quad (\text{B.1})$$

where  $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$ . Given the current estimates  $\mathbf{\Lambda}_{old}$  and  $\mathbf{\Sigma_{eold}}$ , the conditional expectation of the complete data log likelihood function given  $\mathbf{Y}$  is

$$\begin{aligned} E(L_c(\mathbf{\Lambda}, \mathbf{\Sigma_e})|\mathbf{Y}) &= \int \log f(\mathbf{F}, \mathbf{Y}|\mathbf{\Lambda}, \mathbf{\Sigma_e}) f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}}) dF - \frac{n}{2}P_\lambda(\mathbf{\Sigma_e}) \\ &= \int \log(f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}, \mathbf{\Sigma_e}) f(\mathbf{Y}|\mathbf{\Lambda}, \mathbf{\Sigma_e})) f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}}) dF - \frac{n}{2}P_\lambda(\mathbf{\Sigma_e}) \\ &= L_p(\mathbf{\Lambda}, \mathbf{\Sigma_e}) + E \log \left( \frac{f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}, \mathbf{\Sigma_e})}{f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}})} \right) + E \log(f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}})). \end{aligned}$$

Since  $-\log(\cdot)$  is convex, we have

$$-E \log \left( \frac{f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}, \mathbf{\Sigma_e})}{f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}})} \right) \geq -\log E \left( \frac{f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}, \mathbf{\Sigma_e})}{f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}})} \right) = 0$$

by Jensen's inequality. Hence, we get

$$E(L_c(\mathbf{\Lambda}, \mathbf{\Sigma_e})|\mathbf{Y}) \leq L_p(\mathbf{\Lambda}, \mathbf{\Sigma_e}) + E \log(f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}})). \quad (\text{B.2})$$

Given  $\mathbf{\Sigma_{eold}}$  and  $\mathbf{\Lambda}_{old}$ , we update  $\mathbf{\Lambda}$  by minimizing  $-E(L_c(\mathbf{\Lambda}, \mathbf{\Sigma_{eold}})|\mathbf{Y})$ . Thus, we have

$$E(L_c(\mathbf{\Lambda}_{new}, \mathbf{\Sigma_{eold}})|\mathbf{Y}) \geq E(L_c(\mathbf{\Lambda}_{old}, \mathbf{\Sigma_{eold}})|\mathbf{Y}).$$

Then, we update  $\mathbf{\Sigma_e}$  by minimizing  $-E(L_c(\mathbf{\Lambda}_{new}, \mathbf{\Sigma_e})|\mathbf{Y})$ . Since the MM procedure is nonincreasing and the APG algorithm is convergent, we get

$$E(L_c(\mathbf{\Lambda}_{new}, \mathbf{\Sigma_{enew}})|\mathbf{Y}) \geq E(L_c(\mathbf{\Lambda}_{new}, \mathbf{\Sigma_{eold}})|\mathbf{Y}).$$

According to (B.2), the following inequalities hold

$$\begin{aligned}
L_p(\mathbf{\Lambda}_{new}, \mathbf{\Sigma}_{new}) + \mathbb{E} \log(f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma}_{old})) &\geq \mathbb{E}(L_c(\mathbf{\Lambda}_{new}, \mathbf{\Sigma}_{new})|\mathbf{Y}) \\
&\geq \mathbb{E}(L_c(\mathbf{\Lambda}_{old}, \mathbf{\Sigma}_{old})|\mathbf{Y}) \\
&= L_p(\mathbf{\Lambda}_{old}, \mathbf{\Sigma}_{old}) + \mathbb{E} \log(f(\mathbf{F}|\mathbf{Y}, \mathbf{\Lambda}_{old}, \mathbf{\Sigma}_{old})).
\end{aligned}$$

Just dividing the both sides of the above inequalities by  $-n/2$ , the nonincreasing property of the EM plus APG algorithm follows easily.  $\square$

## References

- Bai, J.: Inferential theory for factor models of large dimensions. *Econometrica* **71**(1), 135-171 (2003)
- Bai, J., Liao, Y.: Efficient estimation of approximate factor models via penalized maximum likelihood. *J. Econometrics* **191**(1), 1-18 (2016)
- Bai, J., Ng, S.: Large dimensional factor analysis. *Found. Trends Econom.* **3**(2), 89-163 (2008)
- Bai, J., Shi, S.: Estimating High Dimensional Covariance Matrices and its Applications. *Ann. Econom. Finance* **12**(2), 199-215 (2011)
- Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2**(1), 183-202 (2009)
- Bickel, P., Levina E.: Covariance regularization by thresholding. *Ann. Statist.* **36**(6), 2577-2604 (2008)
- Bien, J., Tibshirani, R.: Sparse estimation of a covariance matrix. *Biometrika* **98**(4), 807-820 (2011)
- Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), 1-122 (2011)

- Chamberlain, G., Rothschild, M.: Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica* **51**(5), 1281-1304 (1983)
- Choi, I.: Efficient estimation of factor models. *Econometric Theory* **28**(02), 274-308 (2012)
- Cui, Y., Leng, C., Sun, D.: Sparse estimation of high-dimensional correlation matrices. *Comput. Statist. Data Anal.* **93**, 390-403 (2016)
- Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**(456), 1348-1360 (2001)
- Fan, J., Liao, Y., Mincheva, M.: Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75**(4), 603-680 (2013)
- Fan, J., Liao, Y., Liu, H.: An overview on the estimation of large covariance and precision matrices. *The Econometrics Journal* **19**(1), 1-32 (2016)
- Forni, M., Hallin, M., Lippi, M., Reichlin, L.: The generalized dynamic-factor model: Identification and estimation. *Rev. Econom. Stat.* **82**(4), 540-554 (2000)
- Geweke, J.: The dynamic factor analysis of economic time series models. Social Systems Research Institute, University of Wisconsin-Madison (1978)
- Hiriart-Urruty, J. B., Lemaréchal, C.: Convex analysis and minimization algorithms II: Advanced theory and bundle methods, vol. 306 of Grundlehren der mathematischen Wissenschaften (1993)
- Hirose, K., Yamamoto, M.: Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Stat. Comput.* **25**(5), 863-875 (2015)
- Lawley, D., Maxwell, A.: Factor analysis as a statistical method, second ed. Butterworths, London (1971)
- Ludvigson, S., Ng, S.: A factor analysis of bond risk premia. In: Ulah, A., Giles, D. (Eds.), *Handbook of Empirical Economics and Finance*, pp. 313-372. Chapman and Hall (2011)

- Ma, S., Xue, L., Zou, H.: Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Comput.* **25**(8), 2172-2198 (2013)
- McLachlan, G., Krishnan, T.: The EM algorithm and extensions. John Wiley & Sons, New York (2007)
- Mulaik, S.A.: Foundations of factor analysis, second ed. CRC press, Boca Raton (2009).
- Ross, S.A.: The arbitrage theory of capital asset pricing. *J. Econom. Theory* **13**(3), 341-360 (1976)
- Rothman, A.J., Bickel, P., Levina, E., Zhu, J.: Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2**, 494-515 (2008)
- Rubin, D.B., Thayer, D.T.: EM algorithms for ML factor analysis. *Psychometrika* **47**(1), 69-76 (1982)
- Stock, J.H., Watson, M.W.: Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97**(460), 1167-1179 (2002)
- Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**(1), 267-288 (1996)
- Xue, L., Ma, S., Zou, H.: Positive-definite  $l_1$ -penalized estimation of large covariance matrices. *J. Amer. Statist. Assoc.* **107**(500), 1480-1491 (2012)
- Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**(2), 894-942 (2010)
- Zou, H.: The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**(476), 1418-1429 (2006)